睡美人文献识别方法研究进展*

■ 宗张建
南京医科大学图书馆 南京 211166

摘要：[**目的/意义**]睡美人文献是一类蕴藏着巨大科学价值的文献。对国内外睡美人文献识别方法的相关研究成果进行总结与梳理，形成比较完整的综述内容，为国内该领域研究提供参考与借鉴。[**方法/过程**]按方法属性将识别方法总结为 4 类，概括各识别方法的基本思路、识别标准、优点及局限性，并参照睡美人文献识别方法的 4 条原则进行对比，指出各识别方法的适用范围。[**结果/结论**]睡美人文献识别方法已由单一逐渐丰富，从主观参数向客观指标演变，从单一引文曲线向多种形式曲线并用发展。未来睡美人文献识别研究可从睡眠深度标准再定义、识别方法再组合上深入。此外，还需重视睡美人文献的定性研究和识别方法的验证性研究，重视睡美人引文价值的测度以及预测性研究。

关键词：睡美人文献 人为参数 曲线拟合 客观指标 数据变换 迟滞承认
分类号：G250
DOI：10.13266/j.issn.0252-3116.2019.16.014

引文是科学对话的一种方式，表明了新知识对原有知识的使用情况。引文的变化反映了学术研究的动态变化，可揭示知识演化、扩散的潜在机制。在文献计量学中，文献的被引次数反映在时间上的曲线称为引文曲线。引文曲线有多种形态，如经典引文曲线、指数增长引文曲线、睡美人引文曲线、双峰引文曲线、波型曲线等[1]。其中，睡美人引文曲线表现为论文发表之后很少引用，但一段时间之后突然被大量引用。荷兰计量学家 A. F. J. Van Raan 巧妙地借用睡美人童话把这一特殊的引文现象命名为睡美人现象（sleeping beauty）[2]。A. F. J. Van Raan 并不是第一个关注到这一现象的学者，相关的表述还有延迟承认（delayed recognition）、超前发现（being ahead of time）、抵制发现（resisted discoveries）、早熟发现（premature discoveries）、孟德尔综合症（Mendel syndrome）等[3]，只是他形象而有趣的命名为科学计量学研究注入了趣味和活力，激发了中外学者更为广泛的研究。叶鹰教授甚至认为：科学睡美人现象与睡美人童话的类比是科学与文学成功交融的一个范例[4]。然而，并不是所有的学者都赞同这一观点。C. R. Sugimoto 等指出：科学追求的是精准，科学术语需要经过严谨地论证，应避免在学术出版物中使用文学隐喻[5]。由此，近年来的一些表述如睡眠文献（sleeping papers）[6]、冬眠文献（hibernator）[7]等就是基于这一观点做出的改变。

睡美人文献形成的要素可归纳为 3 个方面[8-10]：一是科学家特性，包括作者学术资历浅、论文写作能力不足、未与同行充分交流或交流语言障碍等；二是客观环境，包括科学家所处的科研环境、科学共同体、当时的社会经济政治环境等；三是科学发现的本身，包括研究内容的超前性（如研究结果与当前公认的理论不一致、因技术原因无法通过实验扩展为主流认知、跨学科交流失败、研究方法跨领域应用的合法性）、选择错误类型期刊发表论文或期刊影响力低，这也是最重要的。总之，睡美人文献虽稀有，但极富价值，往往与科学研究中的重大发现相关联。因此，实现一定程度的识别，不仅可以使科学的历史学家和社会学家更好地理解科学创新的过程，更有助于缩短科学认知周期、保护重大科学发现和促进科学发展[11]。鉴于此，本文首先对国内外睡美人文献识别方法进行系统梳理归纳，概括各识别方法的基本思路，总结各方法的优点与局限性，其次参照睡美人文献识别方法的 4 条原则进行对比，指出各识别方法的适用范围，最后提出未来的研究方向。

* 本文系南京医科大学哲学社会科学发展专项项目“双一流背景下图书馆学科服务体系的构建及实证研究”（项目编号：2018ZSY006）研究成果之一。
作者简介：宗张建（ORCID：0000-0002-8776-8666），馆员，硕士，E-mail：zzj@njmu.edu.cn。
收稿日期：2018-12-06 修回日期：2019-02-22 本文起止页码：132-142 本文责任编辑：王传清

1 文献搜集方法与相关综述文献

1.1 文献搜集方法

为了保证文献搜集的全面性,笔者采用了 3 步搜集策略:①在 Web of Science(WoS)核心合集 10 个索引子库中进行检索。检索式:TS = (sleeping beauty OR delayed recognition OR being ahead of time OR resisted discoveries OR premature discoveries OR Mendel syndrome)。时间跨度选择所有年份(1900 年-2018 年)。由于检索结果中包含大量的医学、生物学文献。因此又通过 Web of Science 类别——INFORMATION SCIENCE LIBRARY SCIENCE 进一步精炼检索结果,得到 105 条记录。通过阅读文献标题和摘要的方式,剔除掉 30 条与睡美人文献研究内容不相关记录(如文献类型为 Book Review 的文献),最终得到 75 篇相关文献。②在中国知网学术文献总库中进行检索。检索式:主题 =(睡美人 OR sleeping beauty OR 迟滞承认 OR 延迟承认 OR delayed recognition OR 超前发现 OR 抵制发现 OR 早熟发现 OR 孟德尔综合症),并限定在图书情报与数字图书馆领域。检索式中纳入英文关键词,可以检索到在国内期刊发表的英文文献,最终获得相关文献 36 篇。③阅读上述所得文献的参考文献,并通过百度搜索进一步补齐遗漏的 11 篇重要文献,这些文献类型主要是学位论文、会议论文以及 WoS 未收录的研究论文。综上,共计获得相关文献 122 篇。以上文献搜集时间截至 2018 年 11 月 30 日。

1.2 相关综述文献

搜集的文献中,明确为综述文献的有 3 篇:一是张丽华[9]于 2014 年发表的《科学研究中的迟滞承认现象研究进展》,该文梳理了迟滞承认相关概念及界定标准,分析了迟滞承认现象的产生和唤醒机制,在界定标准中作者仅提及平均数识别法和四分位数识别法。二是郭斐等[11]于 2016 年发表的《"睡美人"文献研究综述》,该文聚焦了睡美人文献研究的核心问题,包括识别标准、形成要素、唤醒要素及预测模型,其中识别标准部分描述了平均数识别法、聚类轨迹建模、*B* 指数和四分位数识别法。三是李江[12]于 2016 年发表的《科学中的"睡美人"与"昙花一现"现象评述》,该文讨论了基于平均数标准和基于分位数标准的识别方法,提及 *B* 指数和 G_S 指数。此外,在部分研究论文的文献回顾部分[13],也有对睡美人文献识别方法梳理的探讨。但上述这些讨论或梳理,均缺少必要的归纳对比分析,稍欠全面性和系统性。

目前,睡美人识别方法梳理最为全面的是杜建,在其博士论文《"睡美人"文献的识别方法与唤醒机制研究》中,将睡美人的识别方法总结为 3 类 10 种:曲线拟合法(4 种)、人为参数设定法(2 种)和无参数指标法(4 种)[14]。本文与之不同的是:首先在识别方法选取上,本文舍弃部分识别针对性不足的曲线拟合法,如引文曲线分析框架[1],但同时将最新的研究成果纳入,最终总结为 17 种不同思路的识别方法。其次在识别方法梳理上,将识别方法概括为 4 类,并加强了各方法间的横向比较。这种比较有利于加深各方法间的内在联系以及对识别思路演变规律的认知。最后参照识别方法的 4 条原则逐一检验了 17 种识别方法,指出这些方法的适用范围。

2 睡美人文献识别方法梳理

睡美人文献的识别有定性和定量两种思路。得益于 WoS、Scopus、CNKI 等大型数据库提供的引文数据,睡美人文献的定量研究得以快速发展。本文将视角集中于定量研究,通过对国内外研究文献的梳理,对现有识别方法进行归纳分类,根据方法特性,大致分为人为参数法、曲线拟合法、客观指标法、数据变换法 4 大类,见图 1。下文就依次对这 4 类 17 种识别方法进行梳理。

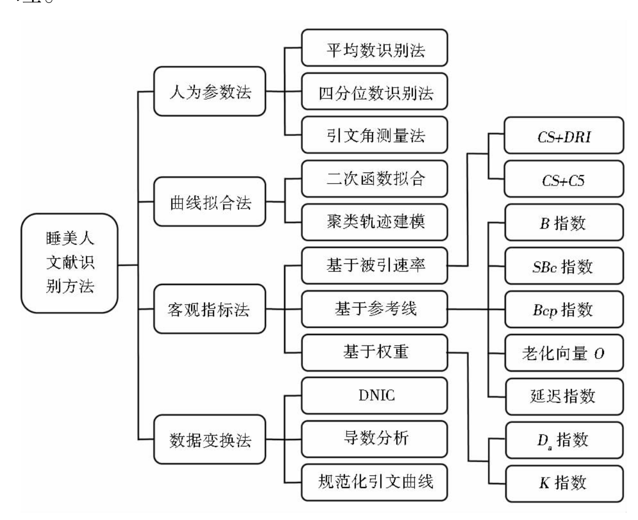


图 1 睡美人文献识别方法框架

2.1 人为参数法

人为参数法就是通过人为定义的阈值来描述睡美人文献的引文特征。此类方法有平均数识别法、四分位数识别法和引文角测量法。

2.1.1 平均数识别法 E. Garfield 于 20 世纪 80 年代末首先引入平均数定义延迟承认,此标准为睡美人文

献的定量研究奠定了基础[15-16]。W. Glänzel 继承了这一思路并开展了更广泛的研究[17-18]。上述研究中，睡美人文献的识别标准始终围绕延迟程度和认可程度展开。延迟程度普遍采用文献发表之初至少 3-5 年的引文窗口，窗口内平均引用次数低于 1 次；认可程度界定标准则差别较大，有超过 50 次、超过 100 次、超过期刊累计影响因子的 10 倍等多种提法。

A. F. J. Van Raan 于 2004 年提出了更清晰明确的睡美人文献界定标准[2]，涉及 3 项指标：①睡眠时长 s（*length of the sleep*）：论文处于沉睡状态的时长；②睡眠深度 C_S（*depth of sleep*）：论文在睡眠时长 s 内，年均至多被引 1 次称为深度睡眠（deep sleep），年均被引 1-2 次称为睡眠（less deep sleep）；③唤醒强度 C_W（*awake intensity*）：紧接睡眠期之后的 4 年唤醒时期（awaking period）内被引用次数（不含自引），唤醒强度要求不低于 20 次。2015 年，A. F. J. Van Raan 又进一步完善了唤醒期 a 和唤醒强度 a_{min}（唤醒期年平均最低引用次数）[19]。通过对不同类型文献集（时间[20]、期刊[21-22]、学科领域[23-24]）的实证研究，A. F. J. Van Raan 的三指标法得到学者的广泛认可。三指标法还为引文角测量法[25-26]、二次函数拟合法[27]等识别法提供了阈值参考，李江等开展的从天才论文（genius work）中识别睡美人文献的标准也借鉴了睡眠深度、唤醒强度等概念[28]。

平均数识别法代表性研究的比较见表 1。此类方法阈值定义主观严格，未考虑学科的差异，虽提高了识别精度，但另一方面导致识别率极低。

表 1　平均数识别法代表性研究综合比较

方法	领域作者	基本思路	识别标准	优点	局限性	实证研究	
						文献集	识别数量
平均数识别法	E. Garfield [15-16]	定义延迟程度和认可程度	高被引论文发表后至少 5 年低被引；初始被引量要求足够低，典型延迟承认论文初始年平均引用频率接近 1 次	简单、直观、易于观察	阈值定义主观、严格	-	5 篇
			论文发表 10 年内，引用次数不超过 10 次；论文发表 20 年时引用次数较之前增加 10 倍			-	20 篇
	W. Glänzel [17-18]		论文发表（a）初始 3 年引用 1 次，或（b）初始 5 年引用最多 2 次；其后论文总被引次数至少达到 100 次			≈45 万	a 标准 77 篇 b 标准 29 篇
			论文发表初始 5 年被引次数极少；随后 15 年获得至少 50 次引用或 10 倍于发表期刊 20 年影响因子累计之和			≈45 万	60 篇
	A. F. J. Van Raan[2]	定义睡眠时长 s、睡眠深度 C_S、唤醒强度 C_W	$s=5\sim10$，$C_S\leq2$，$C_W>20$			≈100 万	359 篇

2.1.2　四分位数识别法　四分位数识别法由 R. Costas 等于 2010 年提出[29]。基本思路是：首先将文献出版后获得 50% 引用次数所需的时间定义为“Year 50%”，其次构建同年同学科所有文献“Year 50%”的分布函数，最后通过分布函数确定 75% 的文献达到 50% 的引用次数所需的时间（P75）。当某篇文献 Year 50% ≥P75，也就是说文献获得一半引用次数所需的时间大于等于 75% 的文献达到一半引用次数所需的时间，该文献则被认为是睡美人文献。

四分位数识别法考虑了文献的整个引文窗口。但此方法识别精度欠佳，按此方法识别出的睡美人文献占文献总量的 25%，不符合睡美人是罕见现象的定义。此外，跨学科文献的学科归属也会给识别带来困难。

2.1.3　引文角测量法　叶鹰等[25-26]于 2017 年提出动态引文角识别睡美人文献的方法。该方法主要步骤是：在文献的年度引文曲线中，把论文发表前一年的时间点定义为零点（0,0），在零点处，时间和引文数均为 0。直线 l 为零点与引文高峰的连线，引文角 β 是直线 l 和时间横轴之间的夹角。计算方法如下：

$$\beta(c,t)=\arctan\left(\frac{c}{t}\right)$$

叶鹰等继续把时间窗口的中点标记为 t_h。令 t_1 为引用前期峰值年份（$t_1<t_h$），引用次数为 c_1，早期引文角为 β_1；t_2 为引用后期峰值年份（$t_2>t_h$），引用次数为 c_2，后期引文角为 β_2，见图 2。直观的，$\beta_1<<\beta_2$ 倾向于产生睡美人文献。设 $t_2-t_1\geq10$。记 t_2-t_1 时段的年均引文为 AC，$t<t_1$ 前 4 年的引文总量为 Ca，$t>t_2$ 后 4 年的引文总量为 Cb。当 $t_2-t_1\geq10$ 时，Cb>20，AC≤2，引文角 $\beta_2>5°$，该文献被认为是睡美人文献。实证研究显示，WoS 数据库 1980 年出版的166 870篇自然科学领域论文中，符合上述标准的睡美人文献有 126 篇[26]。

引文角测量法是一个半经验化测度方法，其重要参考指标 AC、Cb 的阈值参考了三指标法。由于三角函数数值随角度的变化不均匀，比如 tanβ，当 β 角在较

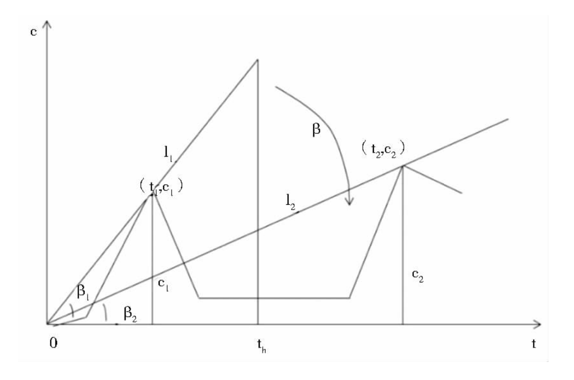


图 2　引文角分析框架[25－26]

小角度变化时,其正切值变化微小;但当 β 角在大角度变化时,其正切值变化显著且迅速。因此,引文角测量法的敏感度不高。

2.2　曲线拟合法

曲线拟合是指选择适当的曲线类型来拟合观测数据,并用拟合的曲线方程分析变量间的关系。利用曲线拟合法识别睡美人文献,有二次函数拟合和聚类轨迹建模两种思路。

2.2.1　二次函数拟合　宋呈玉等[27]提出用二次函数拟合睡美人文献的引用轨迹。基本思路是:一元二次函数表达式为 $y = A + Bx + Cx^2$ ($C \neq 0$),其图像是一条抛物线。睡美人文献的引文曲线类似于开口向上的二次函数图像,对称轴的位置可反映文献发表之初年被引量的大小。依据三指标法中睡美人文献至少沉睡 5 年,二次函数前半部分为睡美人文献沉睡期年度引用量的拟合。研究者利用 OriginPro 工具,对 1998 年－2002 年 WoS 收录的图书情报领域文献进行实证研究。选取被引次数≥20 次的 1 764 篇论文,并限定了 15 年的引文时间窗口。研究发现 $C > 0$ 且对称轴位于 2.5－5.5 之间的拟合曲线较符合,并依此识别出 4 篇睡美人文献。二次函数拟合识别具有操作简便、识别简易的特点,但由于对称轴位置受沉睡时长影响较大,因此对于较长睡眠期的睡美人文献识别效果还需进一步验证。

2.2.2　聚类轨迹建模　S. E. Baumgartner 等[30]将"聚类轨迹建模"(Group-based Trajectory Modeling, GBTM)应用到引文特征分析中,采用的工具为 SAS 数据分析工具加载相关 GBTM 子程序,如 PROC TRAJ。主要步骤为:首先选择零膨胀泊松(ZIP)模型,该分析模型可以观察更多的零值,因而比较适合引文曲线的拟合;其次,利用贝叶斯信息准则(BIC)检验统计量,选择最能代表轨迹间异质性的组数;最后拟合轨迹形状,利用组员平均后验概率(APP)检验模型的充分性。研究者对 6 种期刊和一个学科领域(病毒学)的论文分别进行 GBTM。但只有在 *Journal of the American Society for Information Science* 的 79 篇论文中拟合到有明显特征的睡美人曲线。有趣的是,当拟合组数由 5 组改为 3 或 4 组时,原睡美人组的论文被分配到其他组中。由此可见,该方法的拟合效果与拟合组数密切相关。此外,拟合后组论文数值存在非整数的现象,其原因在于有些论文不能明确地归于某一群体,组论文数是经加权得出。GBTM 存在的诸多局限,表明引文曲线的复杂和多样化。

2.3　客观指标法

睡美人文献的年度引文曲线一般会呈现先抑后扬的状态,即文献发表后几年甚至几十年内被引次数很少,但从某年开始,引用次数快速上升,直至年最大被引量(被引高峰)。通过观察和量化这一过程来识别睡美人文献,有基于被引速率、参考线和权重 3 种思路,合计 9 种方法。

2.3.1　基于被引速率　被引速率(citation speed, *CS*)指标由 J. Wang 提出[31],初用来测度论文被引次数累积的快慢。其计算公式为:$CS = \frac{\sum_{1}^{n-1} C_i / C_n}{n-1}$。其中,$C_i$ 表示第 i 年的累积被引次数,C_n 表示第 n 年的累积被引次数。被引速率的取值在 0－1 之间,被引速率越小,表示论文被引次数累积得越慢。杜建等[32－33]将其引入到睡美人文献识别,认为睡美人文献在引文窗后段的年度被引次数高于前段,具有被引次数累积慢,被引速率小的特点。但实证研究显示,仅以被引速率识别睡美人文献效果并不满意,被引速率低的引文曲线多为直线增长型或指数增长型[33]。

为了弥补这一缺陷,杜建等设计了被引速率与延迟承认指数(*DRI*)[32]或 *C5*[33]组合识别的方法。*DRI* 是论文年度累积被引次数的标准差与被引速率之商,标准差反映论文自发表后被引次数逐年累积程度的差异。睡美人文献年度累积被引次数的标准差大,被引速率小,因而 *DRI* 值较高[32]。*C5* 为人为参数,表示发表之初 5 年内年均被引次数[33]。两种组合识别法综合比较见表 2。

2.3.2　基于参考线　参考线法的基本步骤可概括为 3 步:定义引文曲线、设置参考线和累积参考线与引文曲线之间的距离或面积。根据指标计算方式和参考线设置的差异,此类方法包括 *B* 指数(beauty coefficient)[34]、*SBc* 指数[35]、*Bcp* 指数[36－37]、老化向量 *O*[38] 和延迟指数(*DR*)[39] 5 种,综合比较见表 3。

表 2　基于被引速率的客观指标识别法综合比较

方法	领域作者	基本思路	计算公式	识别标准	优点	局限性	实证研究	
							数据集	识别数量
$CS+DRI$ 组合	杜建等[32]	CS:反映被引次数累积的快慢 标准差:反映论文年度累积被引次数的离散程度	CS 计算方法同上 $DRI=\frac{年度累积引用次数的标准差}{被引速率}$	高被引论文文献中,CS 值越小且 DRI 值越大,睡美人特征越明显	被引速率可观察全部引文曲线,消除论文年龄的影响	方法较为复杂,识别标准不明确,未进行大规模实证检验	学者文献集,21 篇论文	2 篇
$CS+C5$ 组合	杜建等[33]	CS:反映被引次数累积的快慢 $C5$:约束睡眠期的引用次数	CS 计算方法同上 $C5$:发表之初 5 年内年均被引次数	高被引论文且 $CS<0.4$,$C5<5.6$	同上	$C5$ 阈值定义具有人为主观性	期刊文献集,939 篇高被引论文	10 篇

表 3　基于参考线的客观指标识别法综合比较

方法	领域作者	基本思路		计算公式	识别标准	优点	局限性
		定义引文曲线	设置参考线				
B 指数	Q. Ke 等[34]	x 轴:论文年龄 y 轴:论文年度被引次数	论文发表年被引次数点$(0,c_0)$与引文峰值年被引次数点(t_m,t_{ct_m})的连线	$B=\sum_{t=0}^{t_m}\frac{\frac{c_{t_m}-c_0}{t_m}\cdot t+c_0-c_t}{\max\{1,c_t\}}$	数值越大,睡美人特征越明显	客观指标	仅观察部分引文曲线,参考线设置易受引用曲线波动干扰,总被引次数较低睡美人文献的鉴别力不大
SBc 指数	F. Peruzzo[35]	x 轴:论文年龄 y 轴:论文年度累积被引次数	原点(0,0)与论文年度累积被引次数最大点(Δt_m,$y(\Delta t_m)$)的连线	$SBc=\sum_{\Delta t'=0}^{\Delta t_m}\left[\frac{y(\Delta t_m)}{\Delta t_m}\Delta t'-y(\Delta t')\right]$	数值越大,睡美人特征越明显	适合总被引次数较低的文献	指数与文献总被引次数相关性高
Bcp 指数	杜建等[36-37]	x 轴:论文年龄 y 轴:论文年度被引次数累积百分比	论文发表年被引次数点$(0,c_0)$与年度被引次数累积百分比最大点$(t_m,1)$的连线	$Bcp=\sum_{t=0}^{t_m}\frac{1-c_0}{t_m}\cdot t+c_0-c_t$	数值越大,睡美人特征越明显	考虑全部引文曲线,并约束早期引用	尚不明确
老化向量 O	孙建军等[38]	x 轴:论文年龄累积百分比 y 轴:论文年度被引次数累积百分比	原点(0,0)与点(1,1)的连线	$O=(G_S,A^-)$:包含 G_S 和 A^- 为两个参数 $G_S=\begin{cases}1-\frac{2\times[n\times c_1+(n-1)\times c_2+\wedge+c_n]-C}{C\times n}, & C>0\\ 1, & C=0\end{cases}$ A^-:参考线上方与论文累积百分比曲线之间的面积	G_S 值接近 1 且 $A^-=0$	考虑全部引文曲线	文献年龄、所属学科、引文曲线形态对指数有显著影响;无法区分年引用量呈倍数关系的文献;全要素睡美人识别困难[41]
延迟指数(CR)	R. Rousseau[39]	x 轴:论文年龄 y 轴:论文年度累积被引次数	参考线随年份的不同而不同,t 年的参考线为原点(0,0)与$(t,C(t))$的连线	$S(t)=\sum_{n=0}^{t}\left[\frac{C(t)}{t}n-C(n)\right]$ $K(t)=\frac{2}{(t-1)C(t)}S(t)$ $DR(T)=Max_{10\leqslant t\leqslant T}\{0,K(t)\}$	需为当年发表前 1% 高被引论文;且 $DR(T)>0.333$	考虑全部引文曲线	指标计算复杂,需计算每一年的 DR 值;不适合大规模筛选;且未进行大规模实证检验

B 指数,又称美丽系数。该方法直接采用文献的年度引文曲线,定义横坐标 t 为论文年龄,纵坐标 c_t 为论文年度被引次数,设论文发表年被引次数点$(0,c_0)$与引文峰值年被引次数点(t_m,c_{t_m})的连线为参考线,计算参考线年度对应值与年度被引次数的差,并累积该差值与年度被引次数(若年度被引次数为 0,则记为 1)的比值[34]。B 指数对那些特征显著的睡美人文献敏感,当一篇文献睡眠时间越长,睡眠深度越深,唤醒后年度被引次数越高,相应的 B 值就会越大。SBc 指数的思路和 B 指数基本一致,主要不同点在于将纵坐标"论文年度被引次数"修改为"年度累积被引次数"。此方法一定程度上能弥补 B 指数对总被引次数较低的睡美人文献(如总被引次数小于 50 次)识别不足的问题。Bcp 指数则将纵坐标修改为"年度被引次数累积百分比",参考线定义为点$(0,c_0)$与年度被引次数累积百分比最大点$(t_m,1)$的连线,计算参考线年度对应值与引文曲线对应值的差,累积该差值即得到 Bcp 指数。

B 指数、SBc 指数和 Bcp 指数突破了睡眠时间、唤醒强度等人为定义的局限,但指数值并没有明确的分界值区分睡美人文献与"正常"文献。学者通常把指数值最高的前 1% 文献作为候选睡美人文献。杜建等通过实证研究检验了这 3 项指标对睡美人文献的识别

宗张建. 睡美人文献识别方法研究进展[J]. 图书情报工作,2019,63(16):132－142.

效能。在指标与总被引次数的相关程度上,SBc 指数 ＞B 指数＞Bcp 指数;SBc 指数与总被引相关系数高达 0.6,因此认为 SBc 指数不适合用来识别睡美人文献;而 Bcp 指数不仅反映了观察期内论文全部的引文曲线,而且对论文发表之初被引次数的约束效力显著高于 B 指数,因此识别精度优于 B 指数[36]。

G_S 指数由李江等初应用于文献觉醒概率的测算[40]。孙建军等[38]在此基础上提出文献老化向量 $O=(G_S, A^-)$。在老化向量分析框架中,横坐标定义为论文年龄累积百分比,纵坐标定义为被引次数累积百分比。由此可见,横坐标和纵坐标的最大值均为 1。原点、点(1,1)及其到横轴、纵轴的垂线可围成一个面积为 1 的正方形。定义原点(0,0)至(1,1)的连线为参考线。参考线与年度被引次数累积百分比曲线围成一定面积的图形,其中参考线下方靠近横轴的面积定义为正值,用 A^+ 表示;参考线上方靠近纵轴的面积定义为负值,用 A^- 表示。G_S 就是两部分面积算术和的 2 倍,因此 $G_S \in [-1,1]$。当某篇论文的 G_S 值接近 1 且 $A^-=0$,则该论文可认为是睡美人文献。

延迟指数是基于模糊概念提出[39],延迟承认论文必须经历延迟(如 10 年)和认可(如当年发表的前 1% 的高被引论文)两个阶段。在延迟指数分析框架中,横坐标定义为论文年龄,纵坐标定义为年度累积被引次数。对于一篇年龄为 T 年的非零引用文献,其在 t 年的 $S(t)$ 可用如下方法计算:定义年度累积被引次数曲线任意一时点 $n(0 \leqslant n \leqslant t)$ 累积引用次数为 $C(n)$,设原点(0,0)与 $(t, C(t))$ 的连线为参考线,计算参考线年度对应值与年度累积引用次数的差,累积年度差值即可得 $S(t)$。若论文的首次引用发生在第 t 年,$S(t)$ 可获得理论最大值。此时,$S(t)$ 就是原点(0,0)、$(t-1,0)$ 以及 $(t, C(t))$ 三点围成的三角形的面积。$K(t)$ 就是 $S(t)$ 与理论最大值的比值。由此可见,$K(t) \in [-1,1]$。延迟指数 DR 就是当 $t \geqslant 10$ 时最大的 $K(t)$ 值。线性增长曲线的 $DR(T)=0.333$,其值可作为判断是否延迟承认的参考指标。该方法尚未进行大规模实证验证,识别效力还需进一步考察。

2.3.3　基于权重　权重法就是对不同的年度引用次数赋予不同的权重。在睡美人文献识别中,通常将较大的权重给予后期引用,此类方法有 D_a 指数[42]和 K 指数[43]两种,综合比较如表 4 所示:

表 4　基于权重的客观指标识别法综合比较

方法	领域作者	基本思路	计算公式	识别标准	优点	局限性	实证研究	
							数据集	识别数量
D_a 指数	C. Min 等[42]	某年被引频次赋予 i^a 的权重	$D_a=\frac{\sum_1^n i^a \times c_i}{n^a \times C_n}$,其中,$a$ 为调整系数,其值可取 1/3、1/2、2/3、1、2、3、4,c_i 是 i 年的引文数量,C_n 为 n 年中的引用总量	D_a 值接近于 1,或 D_a 值前 1% 文献	可定义 a 值	不利于识别被引次数发生衰退的睡美人	28 769 篇诺贝尔奖获得者论文,总被引次数不低于 19 次	当 $a=1$ 时,D_a 值最大的前 15 篇论文均是睡美人文献
K 指数	A. A. C. Teixeira 等[43]	类似于标准差,某年被引频次赋予时间跨度平方的权重	$K=\left(\frac{\sum_{i=yop}^{yop+N}(i-yop)^2 noc_i}{\sum_{i=yop}^{yop+N} noc_i}\right)^{0.5}/N$,其中,$i$ 指文献的被引年,yop 指出版年,noc_i 是 i 年的引文数量,N 是时间跨度	K 值接近于 1,或 K 值前 1% 文献	可定义分析时间窗口 N	自定义分析时间窗口有一定随意性	5 296 篇社会科学、商业经济学领域"innovation"主题论文,且总被引次数不低于 20 次	当 $N=20$ 时,识别出 8 篇睡美人文献

D_a 指数可定义调整系数 a,a 值可限制引用总次数对识别效果的影响。从识别结果看,D_a 指数识别出的睡美人多呈沉睡时间长,唤醒后上升快速的形态[42]。这应与后期引用次数赋予更多权重有关。但从另一方面看,D_a 指数可能不利于引用次数已发生衰退的睡美人文献的识别。K 指数的特点与 D_a 指数相似,早期的被引次数受到了限制,后期引用次数越多,对 K 值的累积贡献就越大,因而更有利于识别未来高影响力的论文。K 指数算法的不足是,对于同一引文曲线,选用不同的时间跨度 N,计算出的 K 值不同。经实证检验,选取论文发表年至引用上升最快年的时间窗口,计算得出的 K 值最大。后期年引用量放缓或下降,反而会"拖累"K 值的累积,导致 K 值降低。这也正是李秀霞等的实证研究[44],发现其中更符合睡美人文献特征的第 5 篇论文,其 K 值反而不及第 3 篇、第 4 篇的原因。因此,选择合适的时间跨度,对 K 指数尤为重要。对于单篇论文,我们可以观察引文曲线,选取合适时间跨度。但同于一个数据集,统一的时间跨度,可能会造成部分睡美人文献识别疏漏。

2.4　数据变换法

数据变换,就是通过变换将数据转换适合处理和分析的形式。常见的变换方法包括平滑、聚集、数据概化、规范化和属性构造[45]。严格意义上讲,引文数据变换不是一种识别方法,但通过此方法处理后的引文

曲线,结合一些参数指标,也能有较好的识别效果。引文数据变换通常采取规范化和平滑两种形式,涉及文献动态归一化引文影响力(dynamically normalized impact of citations,DNIC)、导数分析和规范化引文曲线3种方法,如表5所示:

表5 DNIC、导数分析法和规范化引文曲线识别法综合比较

识别方法	领域作者	数据变换方式	计算公式	识别标准	优点	局限性	实证研究	
							数据集	识别数量
DNIC	L. Bornmann 等[47]	以同年、同领域、同类型文献的平均引用次数为基准,归一化引文曲线	$\mathrm{DNIC}_{ij}=\frac{C_{ij}}{E_{kj}},\ k=f(i)$ $E_{kj}=\frac{1}{N_{kj}}\sum_{i\mid k=f(i)}C_{ij}$ i、j、k 分别表示文献号、引用年份和领域;C_{ij}为文献 i 在第 j 年的引用次数;E_{kj}表示 k 领域所有文献在第 j 年中的平均引用次数;N_{kj}是 j 年中 k 领域非零引用文献的数量;$k=f(i)$表示给定文献的所属领域	定义时间窗口中点 t_h。将 $t>th$ 时间段曲线峰值记为 $\mathrm{DNIC}_{peak_t>th}$,曲线峰前所有 DNIC_{ij} 值记为 $\mathrm{DNIC}_{b_peak_t}$,$\mathrm{DNIC}_{peak_t>th}>1.6$,且 $\mathrm{DNIC}_{b_peak_t}<0.4$	可避免年代、学科对引文数量的干扰	跨学科文献不易确定所属学科	1980年-1990年的537 1589篇论文	369篇
导数分析法	H. Fang[48]	FIR低通滤波器平滑引文曲线	$dc(0)=c(1)-c(0)$ $dc(t)=(c(t+1)-c(t-1))/2$ $dc(n)=c(n)-c(n-1)$ 其中,n 为论文年龄,c 为 t 年的引用次数	睡眠期 $dc(t)$ 在0附近,且正负值总量相当;当 $dc(t)\geqslant 2$ 睡眠期结束	引入平滑曲线克服波动干扰,可确定多个唤醒期	导数扩大变化的形状,导致唤醒时间后移	-	-
规范化引文曲线	R. Dey[49]	五年移动平均滤波算法平滑曲线,以文献年度引文最大值为基准,规范化引文曲线	将引文曲线最大值定义基数1,其余点依此标准化缩放,因此规范化的数据值都在[0,1]之间	睡眠期≥10年且睡眠期规范化值均<0.2	有利于发现前期引文相对较多的睡美人文献	不能排除"常青树"论文的干扰	计算机领域的178 383篇论文	5 086篇

文献引用频次不仅受文献类型以及所在学科的影响,同时也受文献出版年的影响。WoS数据库近20年收录的文献量超过20世纪100年文献的总和,文献数量的增加可能增加论文被引的概率[46]。通过归一化处理形成的DNIC引文曲线,可以避免学科、出版年和文献类型差异对引文数量的影响。具体讲,DNIC就是一篇文献的年度引用次数与同年、同类型、同领域文献期望被引次数的比值。若 $\mathrm{DNIC}_{ij}=1$,代表文献年度引用与总体平均水平相等。DNIC局限性在于识别睡美人文献的标准有一定主观性,此外跨学科文献不易确定所属领域。

导数分析法将文献的年度被引次数 $c(t)$ 视为一个离散序列,其导数可反映年度引用变化的速率和方向。由导数的含义可知:导数的正或负表示文献引用次数的增加或减少;导数值为零,表明引用次数保持不变。单纯导数分析引文曲线是受限制的,因为当某年前后两年的波动超过4次,则该年的年度导数值就达到了睡眠期结束的标准。采用的平滑曲线可以克服这种细微波动的干扰。导数分析还可以提取睡美人引文曲线的睡眠,上升和下降时期。但导数分析也扩大形变,导致识别的唤醒时间与实际相比后移。

规范化引文曲线识别法以文献年度引文最大值为基准,这种"相对值"的比较有利于识别前期引用次数绝对数较大,但相对后期引用又很小的睡美人文献。此方法的弊端可能是不能排除"常青树"论文(evergreen papers)的干扰。尽管作者逐一检验了5 086篇睡美人文献的原始引用数据,没有发现上述类型论文,但是在其他学科领域,此方法还需充分检验。

3 睡美人文献识别方法的综合比较

李江等提出睡美人文献识别方法的4条原则[41]:①早期被引次数应受到限制。识别方法若能对早期引用进行限制,则有利于睡美人文献识别的准确率;相反,文献早期过多的引用,会使睡美人特征不显著,导致识别出现偏差。②应考虑全部引文曲线。识别方法若仅考虑部分引文史会疏漏其后有价值的信息,不利于一些引文曲线比较特殊的睡美人文献(如全要素睡美人)的识别,考察全部引文历史则有利于文献引用全貌的展示,也有利于睡美人文献唤醒时间的固定。③睡美人文献的唤醒时间应固定,不应随时间变化。④应避免人为参数界定。

从论文早期被引次数是否受限制上看:四分位数识别法通过"Year 50%"限制早期引用;基于被引速率法、*SBc* 指数和延迟指数通过累积被引次数限制早期引用;*Bcp* 指数、老化向量 *O* 通过被引次数累积百分比限制早期引用;D_a 指数和 *K* 指数通过各年引用量赋予

权重的大小限制早期引用。

从考察的引文时间窗口看,平均数识别法、引文角测量法、二次函数拟合、B 指数只考虑了部分引文曲线。具体讲,平均数识别法仅考察了沉睡期至 4 – 5 年唤醒期的引文时间窗口;二次函数曲线拟合因拟合限制仅考察了 15 年的引文时间窗口;B 指数考察了从发表年至年度被引最大值年之间的引文时间窗口;引文角测量法采用了发表年至后期引文峰及之后 4 年的时间窗口。比较特殊的是 K 指数,可自定义分析的时间跨度,但从实证检验来看,采用论文发表年至引用上升最快年的时间窗口有利于得出最大的 K 值。

从能否确定唤醒时间上看,平均数识别法、B 指数、SBc 指数、Bcp 指数和导数分析法均能提示唤醒时间。平均数识别法一般将睡眠期结束后的 4 – 5 年定义唤醒时间。B 指数、SBc 指数和 Bcp 指数确定唤醒时间的思路基本相同:在各自定义的引文曲线中,年度对应点到参考线的距离(垂径)最大时,其指向的年份就是睡美人文献唤醒的年份。导数分析法则将平滑曲线中睡眠期至随后上升期间,年度引用导数大于 2 的年份定义为唤醒时间。

从唤醒时间能否固定上看,平均数识别法和 B 指数确定的唤醒时间有一定缺陷。平均数识别法唤醒时间的确定存在人为主观性。B 指数由于仅将部分引文曲线纳入考察,对于引文曲线含有多个峰值的睡美人文献,不易确定唤醒时间。此外,B 指数还存在唤醒时点的数学意义与实际意义可能不符的问题,即当通过公式计算得出文献在某年被唤醒,但实际上这一年的被引频次为零或处于较低水平。相较于 B 指数,Bcp 指数框架下计算出的唤醒时间更符合实际情况[36]。导数分析法可与其他指数(如 B 指数)结合分析确定唤醒时间。通过平滑曲线分析确定的唤醒时间有不随时间变化、不易受到波动干扰的优点,并且对于多次唤醒的睡美人,可以确定所有的唤醒时间[48]。

从参数是否人为设定上看。人为参数识别法、数据变换识别法涉及人为定义的阈值。曲线拟合识别法的拟合参数需人为设置,且拟合效果也与设置的参数有关。客观指标识别法除了 $CS + C5$ 组合识别法外,均避免了人为参数设置的弊端。其中 B 指数、SBc 指数为绝对值指标,对论文总被引次数的依赖性较高;Bcp 指数、老化向量 O、K 指数和 D_a 指数为相对值指标,Bcp 指数和 D_a 指数可规避指数对被引次数规模的依赖。但需注意的是,上述客观指标在界定睡美人文献时,并没有严格地区分阈值。比较特殊的是基于模糊概念的延迟指数,识别过程参考了线性增长曲线的 $DR(T)$。

综上,当前睡美人文献识别法中,能同时满足上述 4 条原则的仅有 SBc 指数和 Bcp 指数,但 Bcp 指数适用范围更广。从对文献集的要求看,四分位数识别法和 DNIC 对文献集数据的全面性要求高,文献集数据采集不全会对识别结果有重大影响。从方法简易程度上看,平均数识别法、二次函数拟合、Bcp 指数、K 指数和 D_a 指数更有优势。识别方法综合比较及适用范围如表 6 所示:

表 6 睡美人文献识别方法综合比较

识别方法			是否限制早期引用	是否考虑全部引文曲线	能否确定唤醒时间	能否固定唤醒时间	参数是否人为设定	适用文献(集)范围
人为参数法		平均数识别法	×	×	√	×	√	睡眠期年均被引绝对值低,唤醒快速
		四分位数识别法	√	√	×	–	√	需同年、同学科全部文献集合
		引文角测量法	×	×	×	–	√	多峰、振荡或下降型的引文曲线
曲线拟合法		二次函数拟合	不明确	×	×	–	√	浅睡眠且睡眠时间相对较短
		聚类轨迹建模	不明确	√	×	–	√	文献集年龄相同
客观指标法	基于被引速率	$CS + DRI$	√	√	×	–	√	需结合引文曲线,适合小规模数据集
		$CS + C5$	√	√	×	–	半参数	应无限制
	基于参考线	B 指数	×	×	√	×	×	单峰,或有明显的主峰
		SBc 指数	√	√	√	√	×	引用次数小于 50 次
		Bcp 指数	√	√	√	√	×	应无限制
		老化向量 O	√	√	×	–	×	引文未发生衰退或较少衰退
		延迟指数	√	√	×	–	×	应无限制
	基于权重	K 指数	√	可自定义	×	–	×	引文未发生衰退或较少衰退
		D_a 指数	√	√	×	–	×	引文未发生衰退或较少衰退
数据变换法		DNIC	×	√	×	–	√	需同年、同类型、同领域文献集
		导数分析	×	√	√	√	√	可用于引文曲线变化复杂的文献
		规范化引文曲线	×	√	×	–	√	睡眠期年均引用绝对值较高,但相对后期引用较低

4 不足与未来研究展望

总体来看,睡美人文献识别方法已由单一逐渐丰富,从主观参数向客观指标演变,从单一引文曲线向多种形式引文曲线并用发展。但随着研究深入的同时,一些本质问题逐渐显现。

4.1 睡眠深度标准如何定义?

1964 年 C. Dotter 在 *Circulation* 上发表一篇关于动脉硬化闭塞腔内疗法的论文,该论文在最初的 14 年中仅获得 51 次引用,但从 1979 年开始引用激增,在随后的 10 余年中年均被引 50 次以上。1971 年 J. Folkman 在 *New England Journal of Medicine* 上发表了肿瘤新生血管学说,该论文在最初的 23 年中获得 204 次引用,但从 1995 年开始引用激增,至今已获得6 600余次引用。很明显,这两篇医学领域的论文发表初期年均被引分别为 3.64 和 8.86 次,均超过 A. F. J. Van Raan 定义的睡眠深度的标准。但从科学史视角看,C. Dotter 和 J. Folkman 的论文均在当时较长时间内遭受主流学派的忽视或不认同,她们毫无疑问都应属于睡美人文献[50-51]。这种"超标现象"在用客观指数识别出的睡美人中还颇为常见。人们不禁考虑:A. F. J. Van Raan 的睡眠深度标准是不是需重新定义?睡眠深度标准可从 3 个方面考虑:绝对零被引、近似零被引和低被引。绝对零被引是指在考察的时间窗口内未被引用过,近似零被引的被引用频次限定在 1 -2 次,A. F. J. Van Raan 的标准也正基于此,两者阈值明确,没有争议。但低被引还未有共识的定义域。睡美人文献的低被引阈值可能要综合考虑文献的整体引用情况和所在学科特点才能给出,因此需进一步研究。

4.2 现有识别方法如何组合?

通过文献集识别睡美人文献,不同的识别方法识别结果相差较大。杜建等比较了 *B* 指数和 *Bcp* 指数两种方法识别结果,指标排名 top0.1% 的 20 篇论文重合率为 60%[36]。A. A. C. Teixeira 比较了 *K* 指数与三指标法、*B* 指数识别结果,排名 top1% 的 53 篇论文中,*K* 指数与三指标法重合率竟为 0,与 *B* 指数重合率仅为 25%[43]。这种差异一方面与识别方法的特点有关,另一方面也与睡美人引文曲线的形态有关。睡美人文献引文曲线一般包含沉睡、唤醒、高峰、衰减 4 个阶段(部分睡美人引文曲线尚未观测到衰减)。每个阶段又有诸多影响因素,如睡眠时长的长短、睡眠期的引用状态(绝对低被引、昙花一现、相对低被引)、唤醒至引文峰所用时长、唤醒速度(快速、缓慢)、引文峰后状态(振荡、下降)等。复杂的形态增加了识别的难度,单一识别方法不能同时满足各种形态的睡美人曲线,因此可以考虑对现有指标进行组合识别。目前组合多以单一客观指数 + 主观参数认定,如 *K* 指数 + 三指标法[44]、被引速率 + 发表最初 5 年年均被引次数[33]等,但还尚未有多种方法组合识别的研究。多种方法组合识别应达到互为约束、互为补充的效果,那么如何组合现有识别方法?每种组合识别的灵敏度、特异度如何?哪种组合能达到最优的识别效果?这些也需进一步研究。

4.3 重视识别结果的定性研究和识别方法的验证性研究

通过定量研究识别出来的睡美人文献,从严格意义上讲,还应属于潜在的睡美人文献。睡美人文献的最终确定需经科学史或社会学分析等定性方法进行批判性检验[51]。遗憾的是,多数学者对识别结果并未做深入地定性探讨,这是睡美人文献识别研究中不足的一面,因此也造成目前识别方法验证性研究缺失。在机器翻译中,机器翻译软件的水平可通过将机器翻译出来的文本与语料库中的标准译本进行对照来判断。那么,在睡美人文献识别方面,也有必要建立相当于标准译本的睡美人文献识别基线。由此,人们可以对各类识别方法进行实证检验,从而判断各识别方法的正确率、误差率。这对优选识别方法、优化识别组合、提高识别成功度有重要意义。

4.4 重视睡美人文献的引文价值测度

现有睡美人文献的识别方法均基于被引频次这个核心计量指标形成的引文曲线展开。但引用的动机复杂,只通过被引频次并不能完全揭示被引文献对施引文献所贡献的学术价值。完整的引文价值测度包含语法和语义两个层面,前者涉及引用频次和引用位置,后者包含引用类型和引用主题[52]。因此,除引用频次外,可能还需考虑以下问题:施引文献引用了睡美人文献,是引言、背景部分的一般性陈述,还是方法、实验部分的重要参考,或是讨论、结论部分的对比依据?是集成、借鉴的正面引用,还是认同、评述的中性引用,或是商榷、批判的否定引用?睡美人文献与施引文献主题分布的相似度又如何?因此,通过引文价值测度可能能更好地理解睡美人文献沉睡和唤醒现象背后所蕴含的科学技术发展机制。

4.5 重视睡美人文献的预测研究

睡美人现象拓展了文献计量对零被引现象的理解,那些传播与利用状况不佳的文献并非没有价值,它们之中也可能存在潜在"精品"[53]。睡美人文献识别

方法属于回顾性、历史性的识别,但从科学技术促进的视角看,从这些零被引、低被引文献中进行预测性识别才更有意义。目前,关于睡美人文献预测研究成果不多,多为理论研究,可分为两类。一类是通过引用模型预测,如 Q. L. Burrell 的随机模型[54]、李江的心跳图谱框架[40]。但睡美人文献的预测并不仅仅是一个数学建模过程,更多的还需结合文献所蕴藏的科学技术价值进行预测。因此,从文献所包含的内容属性进行预测是另一条路径。一般认为,睡美人文献具有多出自跨学科研究和综合性期刊、多具有潜在技术与应用属性、多为高质量研究的特征,识别变革性研究并追踪其技术转化应用状况,是预测睡美人文献的关键线索[55]。因此,综合此线索并采用一定的预测方法或综合模型,开展前瞻性实证预测研究是未来最有价值的研究方向。

参考文献:

[1] 李江, 姜明利, 李玥婷. 引文曲线的分析框架研究——以诺贝尔奖得主的引文曲线为例[J]. 中国图书馆学报, 2014,40(2):41－49.

[2] VAN RAAN A F J. Sleeping Beauties in science[J]. Scientometrics, 2004,59(3):467－472.

[3] 郭斐, 鄢小燕. 睡美人文献识别方法分析与改进构想[J]. 图书情报工作, 2016,60(8):93－98.

[4] 叶鹰. “睡美人”释义[J]. 中国图书馆学报, 2014,40(2):49.

[5] SUGIMOTO C R, MOSTAFA J. A note of concern and context: on careful use of terminologies[J]. Journal of the Association for Information Science and Technology, 2018,69(3):347－348.

[6] SONG Y, SITU F, ZHU H, et al. To be the Prince to wake up Sleeping Beauty: the rediscovery of the delayed recognition studies[J]. Scientometrics, 2018,117(1):9－24.

[7] HU X J, HU X N, ZHANG Y N, et al. Hibernators, their awakeners and the roles of subsequent authoritative citers[J]. Malaysian journal of library & information science, 2018,23(1):103－113.

[8] TAL D, GORDON A. Sleeping Beauties of political science: the case of AF Bentley[J]. Society, 2017,54(4):355－361.

[9] 张丽华, 张志强. 科学研究中的迟滞承认现象研究进展[J]. 情报杂志, 2014,33(7):97－102.

[10] PARASCHAKIS A. EPA－0345 － “Sleeping beauties” in science [J]. European psychiatry, 2014,29(S1):1.

[11] 郭斐, 鄢小燕. “睡美人”文献研究综述[J]. 图书馆建设, 2016(5):40－45.

[12] 李江. 科学中的“睡美人”与“昙花一现”现象评述[J]. 大学图书馆学报, 2016,34(3):38－43.

[13] 李贺, 解梦凡, 袁翠敏, 等. 用无参数指标 Bcp 识别睡美人文献及其作者动态 h 指数变化规律[J]. 中国图书馆学报, 2018,44(6):75－89.

[14] 杜建. “睡美人”文献的识别方法与唤醒机制研究[D]. 南京: 南京大学, 2017.

[15] GARFIELD E. Delayed recognition in scientific discovery citation frequency analysis aids the search for case histories[J]. Current comments, 1989,23:3－9.

[16] GARFIELD E. More delayed recognition. Part 2. From inhibin to scanning electron microscopy[J]. Current comments, 1990,9:3－9.

[17] GLÄNZEL W, SCHLEMMER B, THIJS B. Better late than never? On the chance to become highly cited only beyond the standard bibliometric time horizon[J]. Scientometrics, 2003,58(3):571－586.

[18] GLÄNZEL W, GARFIELD E. The myth of delayed recognition [J]. The scientist, 2004,18(11):8.

[19] VAN RAAN A F J. Dormitory of physical and engineering sciences: Sleeping beauties may be sleeping innovations[J]. PLOS ONE, 2015,10(10):e139786.

[20] HO Y, HARTLEY J. Highly cited publications in World War II: a bibliometric analysis[J]. Scientometrics, 2017,110(2):1065－1075.

[21] KOZAK M. Current science has its ‘Sleeping Beauties’[J]. Current science, 2013,104(9):1129－1130.

[22] HUANG T, JENG Y, HSU C, et al. Where are the sleeping beauties and princes in educational technology journals? [EB/OL]. [2018－10－18]. https://www.emeraldinsight.com/doi/pdfplus/10.1108/LHT-12-2016-0157.

[23] OHBA N, NAKAO K. Sleeping beauties in ophthalmology[J]. Scientometrics, 2012,93(2):253－264.

[24] ZAVRSNIK J, KOKOL P. Sleeping beauties in pediatrics[J]. Journal of the Medical Library Association, 2016,104(4):313－314.

[25] 张家榕, 曾继城, 叶鹰. 3S 引文现象的特征测度及学术意义——“睡美人”、“时髦女”与“天鹅”综论[J]. 情报学报, 2017,36(12):1241－1246.

[26] YE F Y, BORNMANN L. “Smart girls” versus “sleeping beauties” in the sciences: the identification of instant and delayed recognition by using the citation angle[J]. Journal of the Association for Information Science and Technology, 2018,69(3):359－367.

[27] 宋呈玉, 李秀霞, 谢瑞霞, 等. 基于二次函数曲线拟合的睡美人文献识别研究[J]. 情报杂志, 2018,37(6):119－123,207.

[28] LI J, SHI D. Sleeping beauties in genius work: when were they awakened? [J]. Journal of the Association for Information Science and Technology, 2016,67(2):432－440.

[29] COSTAS R, VAN LEEUWEN T N, VAN RAAN A F J. Is scientific literature subject to a ‘Sell-By-Date’? A general methodology to analyze the ‘durability’ of scientific documents[J]. Journal of the Association for Information Science and Technology, 2010,61(2):329－339.

[30] BAUMGARTNER S E, LEYDESDORFF L. Group-Based Trajectory Modeling (GBTM) of citations in scholarly literature: dynamic qualities of “transient” and “sticky knowledge claims”[J]. Journal of the Association for Information Science and Technology, 2014,65(4):797－811.

[31] WANG J. Citation time window choice for research impact evaluation[J]. Scientometrics, 2013,94(3):851 -872.
[32] 杜建，武夷山. 基于被引速率指标识别睡美人文献及其“王子”——以 2014 年诺贝尔化学奖得主 Stefan Hell 的睡美人文献为例[J]. 情报学报, 2015,34(5):508 -521.
[33] 杜建，武夷山. 睡美人与王子文献的识别方法研究[J]. 图书情报工作, 2015,59(19):84 -92.
[34] KE Q, FERRARA E, RADICCHI F, et al. Defining and identifying sleeping beauties in science[J]. Proceedings of the National Academy of Sciences of the United States of America, 2015,112(24):7426 -7431.
[35] PERUZZO F. Sleeping beauties and the citation dynamics in the network of scientific papers[EB/OL]. [2018 -10 -18]. http://tesi.cab.unipd.it/50039/1/Peruzzo_Fabio.pdf.
[36] 杜建，武夷山. 一个用于识别睡美人文献的新的无参数指标——基于“Science”和“Nature”上睡美人文献的验证[J]. 情报理论与实践, 2017,40(2):19 -25.
[37] DU J, WU Y. A parameter-free index for identifying under-cited sleeping beauties in science[J]. Scientometrics, 2018,116(2):959 -971.
[38] SUN J, MIN C, LI J. A vector for measuring obsolescence of scientific articles[J]. Scientometrics, 2016,107(2):745 -757.
[39] ROUSSEAU R. Delayed recognition: recent developments and a proposal to study this phenomenon as a fuzzy concept[J]. Journal of data and information science, 2018,3(3):1 -3.
[40] LI J, SHI D, ZHAO S X, et al. A study of the “heartbeat spectra” for “sleeping beauties”[J]. Journal of informetrics, 2014,8(3):493 -502.
[41] LI J, YE F Y. Distinguishing sleeping beauties in science[J]. Scientometrics, 2016,108(2):821 -828.
[42] MIN C, SUN J, PEI L, et al. Measuring delayed recognition for papers: uneven weighted summation and total citations[J]. Journal of informetrics, 2016,10(4):1153 -1165.
[43] TEIXEIRA A A C, VIEIRA P C, ABREU A P. Sleeping beauties and their princes in innovation studies[J]. Scientometrics, 2017,110(2):541 -580.
[44] 李秀霞，邵作运，刘超. 基于 K 值算法的图书情报领域“睡美人”文献识别[J]. 图书情报工作, 2017,61(21):114 -122.
[45] 丁刚毅，杨旭，汤海京. 数据科学导论[M]. 北京：北京理工大学出版社, 2017.
[46] ZONG Z, LIU X, FANG H. Sleeping beauties with no prince based on the co-citation criterion[J]. Scientometrics, 2018,117(3):1841 -1852.
[47] BORNMANN L, YE A Y, YE F Y. Identifying “hot papers” and papers with “delayed recognition” in large-scale datasets by using dynamically normalized citation impact scores[J]. Scientometrics, 2018,116(2):655 -674.
[48] FANG H. Analysing the variation tendencies of the numbers of yearly citations for sleeping beauties in science by using derivative analysis[J]. Scientometrics, 2018,115(2):1051 -1070.
[49] DEY R, ROY A, CHAKRABORTY T, et al. Sleeping beauties in computer science: characterization and early identification[J]. Scientometrics, 2017,113(3):1645 -1663.
[50] GORRY P, RAGOUET P. “Sleeping beauty” and her restless sleep: Charles Dotter and the birth of interventional radiology[J]. Scientometrics, 2016,107(2):773 -784.
[51] EL AICHOUCHI A, GORRY P. Delayed recognition of Judah Folkman's hypothesis on tumor angiogenesis: when a Prince awakens a Sleeping Beauty by self-citation[J]. Scientometrics, 2018,116(1):385 -399.
[52] 祝清松. 科技文献引文价值测度的改进方法[J]. 中国科技期刊研究, 2016,27(7):793 -798.
[53] 潘云涛，梁立明，高继平，等. 论文零被引面面观[M]. 北京：科学技术文献出版社, 2018.
[54] BURRELL Q L. Are “Sleeping Beauties” to be expected? [J]. Scientometrics, 2005,65(3):381 -389.
[55] 杜建，武夷山. “睡美人”文献的重要特征、预测线索与政策启示[J]. 科学学研究, 2018,36(11):1938 -1945.

Review on Identification Methods of Sleeping Beauties in Science

Zong Zhangjian
Nanjing Medical University Library, Nanjing 211166

Abstract: [**Purpose/significance**] Sleeping Beauty (SB) refers to a wealth of published literatures with great scientific value. This paper makes a thorough review on the methods to identify SBs at home and abroad, and aims to provide reference for future research in this field. [**Method/process**] The identification methods are divided into four categories. Their theories, criteria, advantages and limitations are summarized. By comparing the Four Rules, the paper points out the application scope of each. [**Result/conclusion**] From subjective parameter to objective indicator, and one citation curve to multiple curve, identification methods of SBs have been enriched. The future study on SBs identification should focus on the redefinition of depth of sleep and the unification of existing methods. In addition, the qualitative and confirmatory research on the predictive value of SBs citation should also be considered.

Keywords: sleeping beauties subjective parameter curve fitting objective indicator data transformation delayed recognition